A geometry-based generic predictor for catalytic and allosteric sites

Simon Mitternacht[*†] and Igor N. Berezovsky[†]

[*]Computational Biology Unit, Bergen Center for Computational Science and
[†]Department of Informatics, University of Bergen, Norway.

Published in *Protein Engineering Design and Selection* 24:405–409 (2011).

Abstract. An important aspect of understanding protein allostery, and of artificial effector design, is the characterization and prediction of substrate and effector binding sites. To find binding sites in allosteric enzymes, many of which are oligomeric with allosteric sites at domain interfaces, we devise a local centrality measure for residue interaction graphs which behaves well for both large and small proteins. The measure is purely structure-based; it has a clear geometrical interpretation and no free parameters. It is not biased towards typically catalytic residues, a property that is crucial when looking for non-catalytic effector sites, which are potent drug targets.

In an era of an exploding number of protein structures, functional site prediction is an important endeavor – to predict the enzymatic function of uncharacterized proteins and to find both biological and artificial allosteric sites. Allosteric mechanisms, as well as the functional motions of proteins in general, utilize the inherent dynamics of the ligand-free protein [1-4]. Therefore, it is important to find sites that have a potential to affect the conformational ensemble and shift the equilibrium towards desired conformations. Ultimately, both biological allosteric sites and predicted novel ones can be used as drug targets.

There are plenty of analysis tools with varying degree of sophistication that analyze surface pockets [5-10], conserved residues [5,11], electrostatics [11] and dynamics [12] to make predictions of ligand binding sites. Artificial sites are not under evolutionary pressure which means that any method based on sequence conservation is unlikely to succeed. Neither is there any reason to believe that they need to have the same chemical properties as catalytic sites. This is why we found it important to develop a minimalistic, purely geometrical, approach for determining residues that belong to natural and latent binding sites. The calculations are simple, quick, essentially parameter-free and independent of amino acid composition and sequence.

Amitai et al. found that the active site of many enzymes can be predicted by ranking exposed residues by closeness centrality in the residue interaction graph (RIG) [13]. The closeness centrality (global closeness) of a node in a graph is defined as the inverse of the average shortest distance to all other nodes. Chea and Livesay found that filtering residues by amino acid identity

improved the predictive ability of global closeness significantly [14]. In a recent study Slama et al. instead use a local network measure, which they weight based on amino acid identity to improve performance further [15]. When identifying residues that take an active part in catalysis, amino acid-based weights help because there are a few amino acids that are clearly over-represented, but when predicting the general binding pocket such filtering is of lesser importance, as we will show.

For a globular domain global closeness essentially indicates how close to the geometrical center of the protein a given residue is. Residues in surface cavities will thus have higher global closeness than the average surface residues, which probably explains most of the predictive ability of the measure. In non-globular geometries this is however no longer true. Figure 1a illustrates what happens in the case of the linear tetramer of tryptophan synthase, where closeness centrality is simply high at the waist of the protein and the correspondence to surface geometry is almost completely lost. We resolve this issue by devising a local centrality measure. Let n_k denote the number of nodes whose shortest distance from a given node is exactly k. The *local closeness* of degree m for a node is then defined as

$$C_m = \sum_{k=1}^{m} \frac{n_k}{k^2}.$$

Local closeness gives a measure of the connectivity of a node, but for finite m only the closest neighbors are considered. In an isotropic network one would expect n_k to scale approximately as k^2, as this would be proportional to the area of a shell with radius k. Weighting by this factor means that the potentially large number of residues at large distances contribute about as much to the local closeness as do the few residues at closer distances. Figure 1b illustrates how local closeness of order 4 (C_4) stays sane in a case where global closeness clearly degenerates. There is a straightforward geometrical rationale for why local closeness can predict ligand binding sites, as illustrated for an idealized 2D network in Figure 1c. The two large circles in the figure represent the nodes covered in the calculation of C_4, and their shade the factor $1/k^2$. Surface nodes in cavities (red circle) will have higher local closeness than at a flat or ridged surface patch (yellow circle) because more nodes are included in the calculation. In addition to the two closeness measures, we will also analyze the local network measure devised by Slama *et al.* [15], which for simplicity will be called S in this manuscript. We reconstructed the parameters of this measure using the most recent versions of the databases (as of May 2010), which should give practically identical parameters to the (unpublished) ones used in that paper.

We will show that local closeness clearly outperforms global closeness as a score for predicting binding sites in protein complexes (Table 1 and Figure 2). Despite the fact that the measure S performs slightly better than local closeness for complexes, local closeness has an obvious advantage: it has a clear geometric interpretation and is not sequence-based, i.e. it can be used to find binding sites that do not have the chemistry typical of catalytic sites. The goal of the follow-

ing analysis is to show this need for a generic structure-based and sequence-independent measure that allows prediction of both catalytic and non-catalytic binding sites. We use three sets of protein structures to show that (i) our measure can identify catalytic residues in single domains to the same precision as other similar measures, (ii) it performs better than global closeness for complexes, and (iii) it is better than the other measures at finding heterotropic (non-catalytic) effector sites in allosteric proteins.

We measure local and global closeness only for surface residues but S score for all residues. To find surface residues we calculate solvent accessible surface areas (ASA) using the program NACCESS; according to Chea and Livesay [14] we use a cutoff-area of 9 $Å^2$ to determine whether a residue is buried or not. Residues are ranked according to each measure, assigning rank 1 to the residue with the highest value of the measure, 2 to the second, etc. For a fair comparison we introduce a version of S without residue weights, but with the same ASA filtering as for the two closeness measures, which we call S_{ASA}.

To compare the performance of local closeness with other measures for single domains we begin by studying the set of 226 domains from the paper by Slama et al. [15]. Each domain corresponds to a different SCOP superfamily and its catalytic residues are annotated in the Catalytic Site Atlas (CSA) [16]. We use global closeness, local closeness, C_m, and the score S to predict catalytic residues, and let CSA annotations define the active sites. For each protein we identify the highest-ranking catalytic residue using either measure (Figure 2a). Since for values of m between 1 and 5, $m = 4$ gave the best performance (Figure S1), the term local closeness will from here on refer to C_4. This value of m effectively means that only residues closer than 30–40 Å are included in the calculation, which roughly corresponds to the length scale of single domains.

Figure 2a compares the performance of the three measures in predicting active site residues. We do both purely geometric calculations only analyzing exposed residues, and calculations where the scores are weighted by residue type for all measures (see caption). Also included in the figure is the corresponding random histogram, calculated exactly for each protein using the hypergeometric distribution with the total number of residues and the number of catalytic residues as parameters. The two closeness measures give virtually identical results in this single domain analysis. The S score performs notably worse without residue weights (S_{ASA}), but when weighted by residue identity (as in its original form) it is on par with the residue-weighted versions of the other two measures. Since only catalytic residues are sought, residue weighting gives a significant improvement. Calculations of global closeness using residue interaction graphs (RIGs) built on all atoms give slightly better results than the equivalent calculations by Chea and Livesay [14] where the RIG was constructed using only C_α contacts (data not shown). Therefore, it is important to use all atom RIGs in predicting binding sites.

The second set of proteins analyzed is based on the non-redundant set from the database Binding MOAD (downloaded on March 15, 2010) [17]. After we reduced sequence identity from

90 % to 40 % using Cd-hit [18], to avoid any biases from over-abundant homologous proteins, there were 2749 proteins with 3529 ligands left. The database provides the correct biological assembly for each protein and indicates which ligands in the PDB structure are biological. More than half of the proteins in this reduced set are complexes. This allows us to compare the different measures for cases where we expect a significant advantage for the local measures. We define all residues that have at least one heavy atom within 4.5 Å of the ligand to be part of the binding site. If there is more than one distinct biological ligand in the structure they are treated independently. If there are several copies of the same ligand, due to symmetry, only the one with the highest ranking binding site is counted (the ranks are however generally very similar for all copies). The top-ranking residue of each site is used to rank the whole site. To be able to make meaningful comparisons between monomers and complexes the ranks are normalized by the number of residues in the protein in Figure 2b. Since we search for the whole binding site, residue weighting, as in the S score, will only have a small effect. For the same reason, the number of target residues is much larger, increasing the chance of getting good ranks for at least one residue by random, compared to the first set. The three measures perform almost identically for monomers, in accordance with the analysis of the first data set, but local closeness and S score give many more correct predictions than global closeness for the larger complexes. Global closeness is not better than random for oligomers consisting of three chains or more.

Since there is no automatic way to determine if a given ligand is allosteric, we finally collected a set of eleven allosteric enzymes that are regulated by ligand binding and that have both native and regulated forms in the PDB (Table I). Six proteins from the allosteric benchmark set fulfill our criteria [19] and five more proteins were found by literature searches, with a preference for heterotropic regulation. Ligand-binding residues were identified in the same way as for the Binding MOAD set. Scores and accessible surface areas were then calculated for an unliganded structure of the same protein. Hence we do not predict binding sites from structures where the ligands are already present as in the previous analysis. Table I shows the results for all eleven proteins using the different scores and two examples of local closeness surfaces are given in Figure 3. The ranks in Table I are normalized to the total number of binding sites in the protein. This is the most natural normalization, but since it is difficult to automate we could not use it in the previous analyses. All proteins in the table have more than one ligand, i.e. substrate, activator, inhibitor, coenzyme, etc. (there is however no crystal structure available for CHK1 with substrate bound). Ligands that bind to the same site are grouped together. Although the number of proteins is small, some trends are clear from this table. In the monomeric proteins (CHK1, protein kinase A and PTP1B) the measures all have comparable performance (with the exception of S not working for protein kinase A). For most of the larger proteins (those with 4 or more binding sites) local closeness and S outperform global closeness, in accordance with previous analyses.

There are three artificial allosteric inhibitor sites in this set, in CHK1, PTP1B and Caspase-7. In CHK1 the effector site ranks well using local closeness and S_{ASA}, but also lies very close to the active site. The inhibitor site in PTP1B, on the other hand, does not rank well with any measure. The tetrameric protease Caspase-7 has a very obvious inhibitor site (see Figure 3a). Hardy et al. also made this observation when they discovered the site, and speculated that there might exist some unknown natural effector that binds here [20]. In general one would expect natural allosteric sites to be easier to find since these have been optimized by evolution.

In protein kinase A the binding of one substrate changes the conformation in a way that makes binding of the second substrate more likely [21]. In tryptophan synthase the activity at one of the two active sites affects the activity at the other [22]. The rest of the proteins have at least one heterotropic effector site, i.e. a site that is only involved in regulation and not catalysis. These sites are particularly interesting because they can have a different chemistry than catalytic sites and represent the type of site that could be targeted by "allosteric drugs". Out of the ten heterotropic effector sites in Table I, six have normalized ranks of 1.5 or less using local closeness, whereas only two sites rank that well using global closeness and S, and three using S_{ASA} (with a rank-threshold of 2.0, S_{ASA} finds five sites). These results confirm that our measure performs well in predicting not only active sites but also effector sites, The number of proteins is not sufficient to say that local closeness is significantly better than S_{ASA} at predicting non-catalytic binding sites, but it has clear advantages in terms of simplicity and transparency.

In conclusion, we have created a purely geometrical measure for predicting biological and artificial binding sites. This measure has obvious advantages over earlier measures: (i) it works for both monomers and oligomers, (ii) it is parameter-free, (iii) it has a clear and intuitive geometrical interpretation, and (iv) it does not use any information about sequence or amino acid composition. The latter is especially important for predicting artificial binding sites in drug design, as these need not have the same chemical composition as catalytic sites. For finding catalytic residues in single domains the three measures analyzed here (and their variants) are more or less equivalent. For complexes the two local measures, local closeness and S, are clearly better than global closeness. However, the score by Slama et al. [15] has an obfuscated geometrical meaning, and uses residue weighting in its original form, which lowers the chance of finding non-catalytic binding sites. Lastly, in the small set of allosteric proteins we see that effector sites can be predicted with an accuracy comparable to that for active sites, and that local closeness performs better than the other scores for the non-catalytic effector sites.

Acknowledgments

Financial support from the Norwegian Research Council via Functional Genomics Programme (FUGE II) is gratefully acknowledged.

References

1. Cui Q, Karplus M (2008) Allostery and cooperativity revisited. Protein Sci 17: 1295-1307.
2. Weber G (1972) Ligand binding and internal equilibria in proteins. Biochemistry 11: 864-878.
3. Changeux JP, Edelstein SJ (2005) Allosteric mechanisms of signal transduction. Science 308: 1424-1428.
4. Swain JF, Gierasch LM (2006) The changing landscape of protein allostery. Curr Opin Struct Biol 16: 102-108.
5. Glaser F, Morris RJ, Najmanovich RJ, Laskowski RA, Thornton JM (2006) A method for localizing ligand binding pockets in protein structures. Proteins 62: 479-488.
6. Weisel M, Proschak E, Schneider G (2007) PocketPicker: analysis of ligand binding-sites with shape descriptors. Chem Cent J 1: 7.
7. Xie L, Bourne PE (2007) A robust and efficient algorithm for the shape description of protein structures and its application in predicting ligand binding sites. BMC Bioinformatics 8 Suppl 4: S9.
8. Laurie AT, Jackson RM (2006) Methods for the prediction of protein-ligand binding sites for structure-based drug design and virtual ligand screening. Curr Protein Pept Sci 7: 395-406.
9. Tripathi A, Fornabaio M, Kellogg GE, Gupton JT, Gewirtz DA, et al. (2008) Docking and hydropathic scoring of polysubstituted pyrrole compounds with antitubulin activity. Bioorg Med Chem 16: 2235-2242.
10. Kawabata T, Go N (2007) Detection of pockets on protein surfaces using small and large probe spheres to find putative ligand binding sites. Proteins 68: 516-529.
11. Bray T, Chan P, Bougouffa S, Greaves R, Doig AJ, et al. (2009) SitesIdentify: a protein functional site prediction tool. BMC Bioinformatics 10: 379.
12. Ming D, Cohn JD, Wall ME (2008) Fast dynamics perturbation analysis for prediction of protein functional sites. BMC Struct Biol 8: 5.
13. Amitai G, Shemesh A, Sitbon E, Shklar M, Netanely D, et al. (2004) Network analysis of protein structures identifies functional residues. J Mol Biol 344: 1135-1146.
14. Chea E, Livesay DR (2007) How accurate and statistically robust are catalytic site predictions based on closeness centrality? BMC Bioinformatics 8: 153.
15. Slama P, Filippis I, Lappe M (2008) Detection of protein catalytic residues at high precision using local network properties. BMC Bioinformatics 9: 517.
16. Porter CT, Bartlett GJ, Thornton JM (2004) The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. Nucleic Acids Res 32: D129-133.
17. Hu L, Benson ML, Smith RD, Lerner MG, Carlson HA (2005) Binding MOAD (Mother Of All Databases). Proteins 60: 333-340.
18. Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22: 1658-1659.
19. Daily MD, Gray JJ (2007) Local motions in a benchmark of allosteric proteins. Proteins 67: 385-399.
20. Hardy JA, Lam J, Nguyen JT, O'Brien T, Wells JA (2004) Discovery of an allosteric site in the caspases. Proc Natl Acad Sci U S A 101: 12461-12466.
21. Masterson LR, Mascioni A, Traaseth NJ, Taylor SS, Veglia G (2008) Allosteric cooperativity in protein kinase A. Proc Natl Acad Sci U S A 105: 506-511.
22. Barends TR, Dunn MF, Schlichting I (2008) Tryptophan synthase, an allosteric molecular factory. Curr Opin Chem Biol 12: 593-600.

Tables

Table I: Normalized ranks of ligand binding sites in allosteric proteins. The absolute ranks can be obtained by multiplying the ranks by the number of sites. There is no structure available for CHK1 with substrate bound. Non-catalytic binding sites are marked in bold face.

			Ranks normalized by number of sites			
Protein/PDB entries used	Sites	Ligand-name in PDB	LC	GC	S	S_ASA
Anthranilate synthase	6	**TRP (inhibitor)**	4.5	12.8	7.5	2
1i7s, 1i7q		BEZ/PYR (substrate)	2.3	1.8	4.2	1
		ILG (substrate)	4.2	17.7	2.5	5
ATCase	12	PAL (substrate analog)	0.1	0.3	0.1	0.1
1d09, 1rac, 3d7s, 7at1		**ATP/CTP (effector)**	54.6	137.4	40.1	53.8
boGDH	18	**ADP (activator)**	1.2	8.3	0.9	0.9
1nr7, 1nqt, 1hwz		**GTP (inhibitor)**	10.8	24.3	49.6	26
		GLU/NDP (substrate)	1.3	13.9	0.9	0.5
Caspase-7	4	**NXN (inhibitor)**	1	0.5	2.75	0.5
3ibf, 3ibc, 1shj		Peptide substrate	3.5	7.5	1.5	5.75
CHK1 (1ia8, 3jvs)	2	**AGY (inhibitor)**	1.5	2.5	3.5	0.5
GlcN6P deaminase	12	AGP (substrate)	2.3	11.8	1.4	0.7
1cd5, 1hor, 1hot		**16G (activator)**	0.25	4.8	10	3.7
Phosphofructokinase	8	**PGA/ADP1 (effector)**	0.1	3.1	3.1	1.6
3pfk, 4pfk, 6pfk		F6P/ADP2 (substrate)	5.6	7.1	20.1	7.8
Protein Kinase A	1	ATP (substrate)	4	1	13	4
1j3h, 1atp		Peptide substrate	3	3	13	4
PTP1B	2	BPM (substrate)	0.5	0.5	1	0.5
2hnp, 1aax, 1t49		**892 (inhibitor)**	16.5	12	28	15.5
Threonine synthase	4	**SAM (activator)**	1.25	0.25	0.75	0.25
1e5x, 2c2b, 2c2g		LLP (coenzyme)	1.75	7	3.75	2.25
Tryptophan synthase	4	G3H/IDM (substrate 1)	9.25	91.25	0.5	0.25
1bks, 3cep		PLP/SRI/IDM (substrate 2)	0.75	0.25	2	0.75

Figures

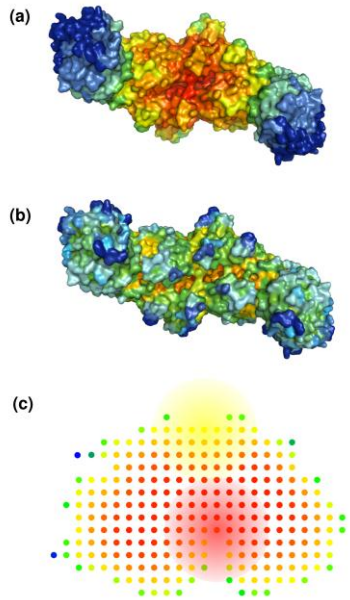

Figure 1: Global (a) and local (b) closeness for tryptophan synthase (PDB code 1bks). To calculate the network measures we generate a residue interaction graph where each residue is a node. We define two residues to be in contact if any of their atoms are within a distance of 5 Å. Hydrogens are added to structures where needed. The programs used for generating and analyzing the RIGs were written in C++ using the Boost Graph Library. Panel (c) illustrates how local closeness can be interpreted geometrically for an idealized RIG (see main text).

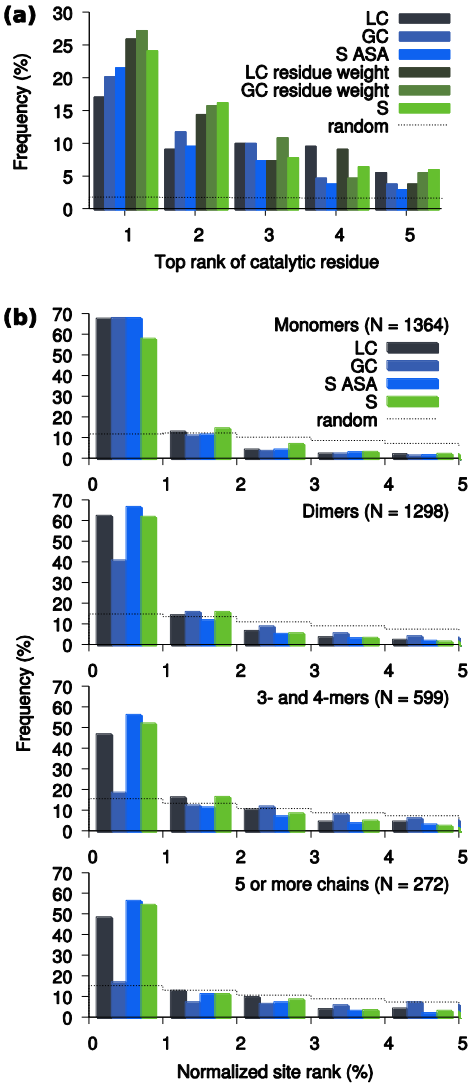

Figure 2: Comparison of local closeness (LC), global closeness (GC) and the score S by Slama et al. (2008). Blue bars indicate purely geometrical measurements where only exposed residues are ranked; the factor weighting different residue types is thus not included the version of S labeled S_{ASA}. Green bars indicate measurements where no ASA screening has been done, but the residue-weight factor from S has been used to weight the scores. The value of the parameter k_{type} (see Slama et al. 2008) is set to 2.8 for LC and 1.7 for GC. Panel (a) shows predictions of catalytic residues for the 226 proteins with CSA annotation and (b) the 2749 proteins derived from Binding MOAD. The bars indicate the distribution of top ransk for catalytic residues in (a) and normalized ranks for binding site residues in (b). This means that the more the distribution is shifted to the left, the better the predictive power of the measure.

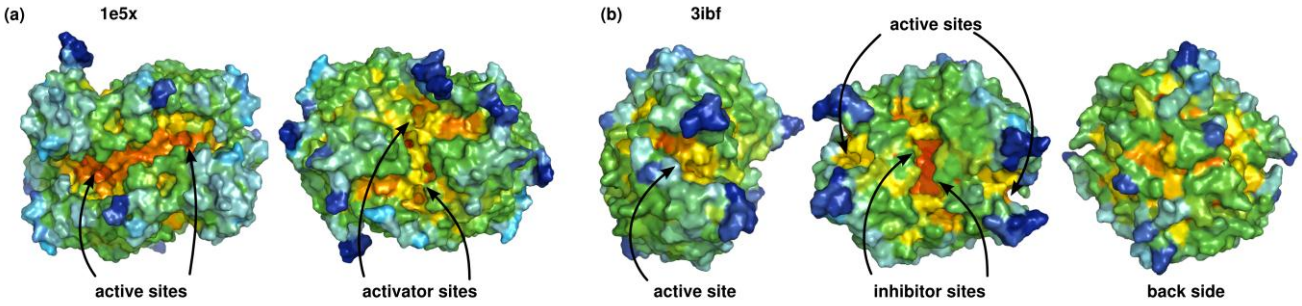

Figure 3: Examples of local closeness surfaces for (a) threonine synthase and (b) Caspase-7. The two activator sites in threonine synthase are buried in internal cavities, one of the openings can be seen as a small red hole near the arrows for each site.

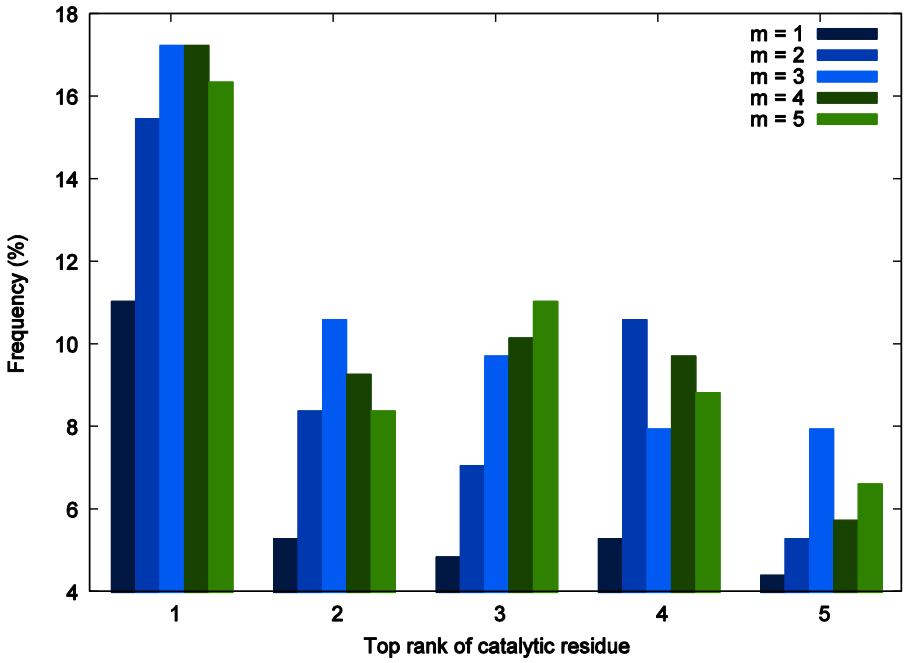

Figure S1: Selection of the parameter m in C_m. The histograms are generated the same way as those in Figure 2a for different values of m. The higher the sum of the bars, the better the measure. The values m = 3 and 4 are more or less equivalent. We chose to use m = 4.